Gaussian processes for classification

Manel Martínez-Ramón

ECE, UNM

October 2019

Generative and discriminative approaches

- In classification, we assume that a joint probability distribution $p(\mathbf{x}, y)$ exists.
- Then, by the Bayes rule we can decompose this probability in

$$p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y) = p(y|\mathbf{x})p(\mathbf{x})$$

The first decomposition is the generative approach, where $p(y\mathbf{x})$ is the class conditional distribution and

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$$

where $p(\mathbf{x}) = \sum_c p(\mathbf{x}|(C)_c)p(\mathcal{C}_c)$ and $\mathcal{C}_c$ is each one of the possible class distributions.

Generative and discriminative approaches

- The discriminative approach is intended to model $p(y|\mathbf{x})$ in a direct way without making inference on $p(\mathbf{x}|y)$.
- Both methods, nevertheless, need probabilistic models. For the generative model, one can use a Gaussian class conditional model

$$p(\mathbf{x}|\mathcal{C}_c) = \mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c).$$

- For the discriminative case, we can establish an activation or response function bounded between 1 and 1. The linear Logistic Regression is often used

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + exp(-\mathbf{x}^\top \mathbf{w})}$$

Generative and discriminative approaches

- For the logistic function, if $\mathbf{x}^\top \mathbf{w}$ is positive, then the probability model tends to 1 when its absolute value increases, and when it is negative, the probability tends to zero if the absolute value increases.
- The Probit function is also used to model a discriminative distribution as

$$p(y = 1|\mathbf{x}) = \int_{-\infty}^{\mathbf{x}^\top \mathbf{w}} \mathcal{N}(\tau, 1) d\tau$$

Generative and discriminative approaches

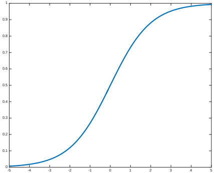
- Discriminative approaches are easy to implement. Generative approaches are hard, and the complexity increases with the number of data
- Generative approaches are able to deal with situations where some labels are missing (semisupervised) or where there are no labelling (unsupervised), while discriminative approaches cannot.
- The approach taken by Gaussian processes are nevertheless discriminative.

Linear classification

- The likelihood of a label y is defined as

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^\top \mathbf{w})$$

where σ is any sigmoidal function.



- It is easy to see that for the logistic function we have

$$\log\left(\frac{p(y = 1|\mathbf{x}, \mathbf{w})}{p(y = -1|\mathbf{x}, \mathbf{w})}\right) = \mathbf{x}^\top \mathbf{w}$$

The quotient is the so called logit function.

Posterior

- Assume a dataset $\{\mathbf{x}_1, y_1 \cdots \mathbf{x}_N, y_N\}$ and use a Gaussian prior $\mathcal{N}(\mathbf{0}, \Sigma_p)$ for the parameters $\mathbf{w}$. Then, we can compute a posterior for the parameters of the form

$$\log p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = -\frac{1}{2}\mathbf{w}^\top \Sigma_p^{-1}\mathbf{w} + \sum_{n=1}^{N} \log \sigma(y_n f_n)$$

  where $f_n = \mathbf{x}_n^\top \mathbf{w}$ and where the constant (normalization) terms have been omitted.

- Note that the first term is a quadratic form, and the second one is monotonic, so the posterior has a single maximum.

Predictive distribution

- The posterior allows us to compute the predictive distribution for a test data

$$p(y=1|\mathbf{x}^*,\mathbf{X},\mathbf{y})=\int p(y=1|\mathbf{x}^*,\mathbf{w})p(\mathbf{w}|\mathbf{X},\mathbf{y})d\mathbf{w}$$

Gaussian process model for classification

- Assume a prior over the latent function $f(\mathbf{x})$ and then a logistic function over this prior.

$$\pi(\mathbf{x}) = p(y = 1|\mathbf{x}) = \sigma(f(\mathbf{x}))$$

- A predictive distribution can then be computed as it is done in regression as

$$p(f(\mathbf{x}^*)|\mathbf{X}, \mathbf{y}, \mathbf{x}^*) = \int p(f(\mathbf{x}^*)|\mathbf{X}, \mathbf{x}^*, \mathbf{f})p(\mathbf{f}|\mathbf{X}, \mathbf{y})d\mathbf{f}$$

where $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ is the posterior of the latent function over the data.

Gaussian process model for classification

- If we use this distribution to construct the posterior over the latent $f(\mathbf{x}^*)$ we produce a probabilistic prediction

$$p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}^*) = \int \sigma(f(\mathbf{x}^*))p(f(\mathbf{x}^*)|\mathbf{X}, \mathbf{y}, \mathbf{x}^*)df(\mathbf{x}^*)$$

- These two integrals are analytically intractable, so we need analytical approximations of integrals or Monte Carlo sampling.
- In GP two approaches are used: The Laplace approximation and the Expectation Propagation (EP) method.

Laplace approximation

- The method is based on a Gaussian approximation of the posterior probability $p(\mathbf{f}|\mathbf{X},\mathbf{y})$.
- The approximation is obtained based on a second order Taylor expansion of the posterior:

$$q(\mathbf{f}|\mathbf{X},\mathbf{y}) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}},\mathbf{A}^{-1}) \propto \exp\left(-\frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^\top \mathbf{A}(\mathbf{f}-\hat{\mathbf{f}})\right)$$

where

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f}|\mathbf{X},\mathbf{y})$$

$$\mathbf{A} = -\nabla\nabla \log p(\mathbf{f}|\mathbf{X},\mathbf{y})|_{\mathbf{f}=\hat{\mathbf{f}}}$$

Laplace approximation: posterior

- By the Bayes' rule

$$p(\mathbf{f}|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})} \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})$$

Taking logarithms

$$\log p(\mathbf{f}|\mathbf{X},\mathbf{y}) \propto \boldsymbol{\Psi}(\mathbf{f}) = \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|\mathbf{X})$$

- The second term has been computed for regression:

$$\boldsymbol{\Psi}(\mathbf{f}) = \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^\top\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi$$

Laplace approximation: posterior

- The first goal here is to find the maximum of this posterior. We use the Newton's method, which needs the gradient and the Hessian:

$$\nabla \boldsymbol{\Psi}(\mathbf{f}) = \nabla \log p(\mathbf{y}|\mathbf{f}) + \mathbf{K}^{-1}\mathbf{f}$$

$$\nabla\nabla \boldsymbol{\Psi}(\mathbf{f}) = \nabla\nabla \log p(\mathbf{y}|\mathbf{f}) + \mathbf{K}^{-1}$$

  where the gradient and the Hessian of $\log p(\mathbf{y}|\mathbf{f})$ depend on the model that we use for this distribution (see textbook, table pag. 43).

Laplace approximation: posterior

- The gradient has to be nulled:

  $$\nabla \log p(\mathbf{y}|\mathbf{f}) + \mathbf{K}^{-1}\mathbf{f} = \mathbf{0}$$

  with which

  $$\hat{\mathbf{f}} = \mathbf{K}\nabla \log p(\mathbf{y}|\hat{\mathbf{f}})$$

  Since this equation is transcendental, we can only find $\hat{\mathbf{f}}$ iteratively as

  $$\mathbf{f}^{k+1} = \mathbf{f}^{k} - (\nabla\nabla\boldsymbol{\Psi})^{-1}\nabla\boldsymbol{\Psi}$$

- We define $\mathbf{W} = -\nabla\nabla \log p(\mathbf{y}|\mathbf{f})$, and then the posterior can be approximated by a Gaussian

  $$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \approx q(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\hat{\mathbf{f}}, (\mathbf{K}^{-1} + \mathbf{W})^{-1})$$

Laplace approximation: prediction

- The goal of the prediction is to find the expectation of the predictive distribution. SInce the approximation is Gaussian, we can use the expression obtined for regression

$$\mathbb{E}[f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*] = \mathbf{k}(\mathbf{x}_*)\mathbf{K}^{-1}\hat{\mathbf{f}}$$

and since

$$\hat{\mathbf{f}} = \mathbf{K}\nabla \log p(\mathbf{y}|\hat{\mathbf{f}})$$

then

$$\mathbb{E}[f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*] = \mathbf{k}(\mathbf{x}_*)\nabla \log p(\mathbf{y}|\hat{\mathbf{f}})$$

Laplace approximation: prediction

- The variance can also be computed by using the expression obtained for regression

$$\mathbb{V}(f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) = k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top(\mathbf{K}+\mathbf{W}^{-1})^{-1}\mathbf{k}_*$$

- Approximations of multiclass classifier exist by using the Softmax activation

$$p(y_i^c|\mathbf{f}_i) = \frac{\exp(f_i^c)}{\sum_c \exp(f_i^c)}$$

Laplace approximation: marginal likelihood

- It is useful to compute an approximatio for the marginal likelihood:

$$\log q(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = -\frac{1}{2}\hat{\mathbf{f}}^\top \mathbf{K}^{-1}\hat{\mathbf{f}} + \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\log|\mathbf{B}|$$

  with $\mathbf{B} = |\mathbf{I} + \mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}|$

- It is necessary in order to find optimal values for the hyperparameters (see Rasmussen, Ch. 5)

Expectation-Propagation

- The EP procedure is used as an alternative way of estimating the parameters of the GP for classification.
- It is based on the factorization of the posterior distribution over the latent variables $\mathbf{f}$:

$$p(\mathbf{f}|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{f}|\mathbf{X})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y}|\mathbf{X})} = \frac{1}{\mathbf{Z}}p(\mathbf{f}|\mathbf{X})\prod_{i=1}^{N} p(y_i|f_i)$$

where $\mathbf{Z} = p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} p(y_i|f_i) d\mathbf{f}$

Expectation-Propagation

- Here, the likelihood over labels y_i is modelled using the probit function

$$p(y_i|f_i) = \Phi(f_i y_i)$$

But this makes the posterior intractable, so we approximate it by a Gaussian

$$p(y_i|f_i) \approx t_i(f_i|\tilde{\mathbf{Z}}_i, \tilde{\boldsymbol{\mu}}_i, \tilde{\sigma}_i^2) = \tilde{\mathbf{Z}}_i \mathcal{N}(f_i, \tilde{\boldsymbol{\mu}}_i, \tilde{\sigma}_i^2)$$

Expectation-Propagation

- With this approximation, the product of (local) likelihoods becomes

$$\prod_{i=1}^{N} t_i(f_i|\tilde{\mathbf{Z}}_i, \tilde{\boldsymbol{\mu}}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \prod_i \tilde{\mathbf{Z}}_i$$

  Where the mean is a vector of all means and the covariance matrix is a diagonal containing all variances $\tilde{\sigma}_i^2$

Expectation-Propagation

- With that, the posterior $p(\mathbf{f}|\mathbf{X},\mathbf{y})$ is approximated as

$$q(\mathbf{f}|\mathbf{X},\mathbf{y}) = \frac{1}{\mathbf{Z}_{EP}} p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} t_i(f_i|\tilde{\mathbf{Z}}_i, \tilde{\boldsymbol{\mu}}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

  which is a Gaussian with parameters $\boldsymbol{\mu} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma} = (\mathbf{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1}$
- The EP chooses then the parameters of the local likelihoods in a sequential way.

Expectation-Propagation

- The algorithm sequentially approximates each t_i by using the current approximations for all other variables. Each t_i contains the following terms:
  1. The prior $p(\mathbf{f}|\mathbf{X})$
  2. The rest of likelihoods t_j
  3. The exact likelihood $p(y_i|f_i) = \Phi(y_i f_i)$
- To this end, we construct the *cavity function*

$$q_{-i}(f_i) \propto \int p(\mathbf{f}|\mathbf{X}) \prod_{j \neq i} t_i(f_i|\tilde{\mathbf{Z}}_i, \tilde{\boldsymbol{\mu}}_i, \tilde{\sigma}_i^2) df_j$$

Expectation-Propagation

- This cavity function is another Gaussian

$$q_i(f_i) = \mathcal{N}(f_i|\mu_{-i}, \sigma{-}i^2) \tag{1}$$

with $\mu_{-i} = \sigma_{-i}^2(\sigma_i^{-2}\mu_i - \tilde{\sigma}_i^{-2}\tilde{\mu}_i)$ and $\sigma_{-i}^2 = (\sigma_i^2 - \tilde{\sigma}_i^{-2})^{-1}$

- The EP algorithm can then be summrized as
  1. computing the marginal for sample i

$$\hat{q}(f_i) \triangleq \hat{Z}_i\mathcal{N}(\hat{\mu}_i u, \hat{\sigma}_i^2) \approx q_{-i}(f_i)p(y_i|f_i)$$

  2. Find the parameters $\tilde{\mu}_i$,$\tilde{\sigma}_i$ of the approximate likelihood that minimize the KL divergence with respect the true likelihood.

We must iterate this procedure for all samples.

Expectation-Propagation

The values of the parameters for each iteration are:

$$\hat{Z}_i = \Phi(z_i) \qquad \hat{\mu}_i = \mu_{-i} + \frac{y_i {\sigma^2}_{-i} \mathcal{N}(z_i)}{\Phi(z_i)\sqrt{1 + {\sigma^2}_{-1}}}$$

$$\hat{\sigma}_i^2 = {\sigma^2}_{-1} - \frac{{\sigma^4}_{-i} \mathcal{N}(z_i)}{(1 + {\sigma^2}_{-i})\Phi(z_i)} \left(z_i + \frac{\mathcal{N}(z_i)}{\Phi(z_i)} \right) \qquad \text{with} \quad z_i = \frac{y_i \mu_{-i}}{sqrt 1 + {\sigma^2}_{-1}}$$

Expectation-Propagation

Then, the moments of the approximate posterior are computed as

$$\tilde{\mu}_i = \tilde{\sigma}_i^2(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_{-i}^{-2}\mu_{-i}), \quad \tilde{\sigma}_i^2 = (\hat{\sigma}_i^{-2} - \sigma_{-i}^{-2})^{-1}$$

$$\tilde{Z}_i = \hat{Z}_i\sqrt{2\pi}\sqrt{\sigma_{-i}^2 + \tilde{\sigma}_i^2}\exp\left(\frac{1}{2}(\mu_{-i} - \tilde{\mu}_i)^2/(\sigma_{-i}^2 + \tilde{\sigma}_i^2)\right)$$

Expectation-Propagation

The predictive mean and variance are:

$$\mathbb{E}[f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*] = \mathbf{k}_* \left(\mathbf{K}+\tilde{\boldsymbol{\Sigma}}\right)^{-1} \tilde{\boldsymbol{\mu}}$$

$$\mathbb{V}[f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*] = k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top \left(\mathbf{K}+\tilde{\boldsymbol{\Sigma}}\right)^{-1} \mathbf{k}_*$$

And the predictive distribution for the target is

$$q(y_*=1|\mathbf{X},\mathbf{y},\mathbf{x}_*) = \mathbb{E}[\pi_*|\mathbf{X},\mathbf{y},\mathbf{x}_*] = \int \Phi(f_*)q(f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*)df_*$$

that becomes

$$q(y_*=1|\mathbf{X},\mathbf{y},\mathbf{x}_*) = \Phi\left(\frac{\mathbb{E}[f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*]}{\sqrt{1+\mathbb{V}[f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*]}}\right)$$

Algorithm

- Input: $\mathbf{K}$ and y
- Initialize parameters of slide 21 to zero.
- Repeat until convergence
  1. Compute parameters of eq. (1)
  2. Compute moments of slide 24
  3. Update parameters of slide 25
  4. Update Σ and $\boldsymbol{\mu}$